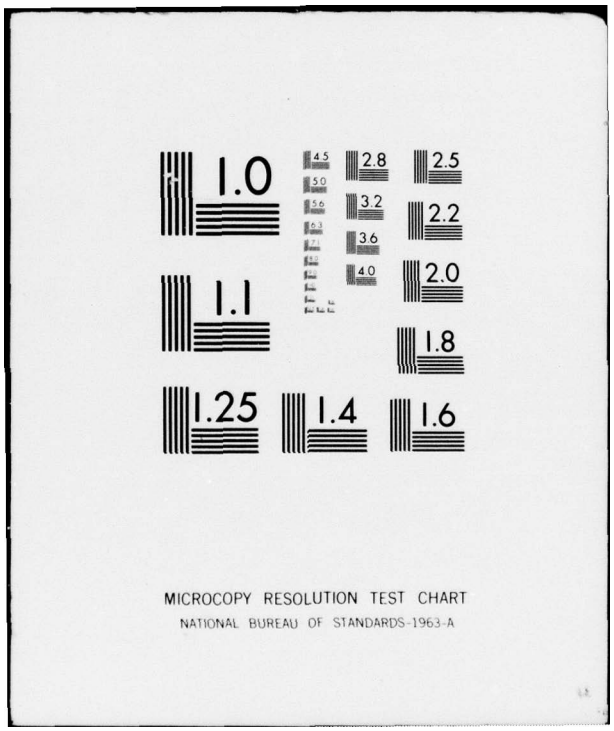
AD-A040 332 SPEECH COMMUNICATIONS RESEARCH LAB INC SANTA BARBARA CALIF F/G 9/4
ACOUSTIC/LINGUISTIC ASPECTS OF AUTOMATIC SPEECH RECOGNITION.(U)
APR 77 D J BROAD, L L PFEIFER F44620-74-C-0034

UNCLASSIFIED AFOSR-TR-77-0631 NL

| OF |
AD
A040332

END

DATE
FILMED
7-77


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

INTERIM PROGRESS REPORT

OF RESEARCH ON

ACOUSTIC/LINGUISTIC ASPECTS OF AUTOMATIC SPEECH RECOGNITION

David J. Broad, Ph.D.

Larry L. Pfeifer, Ph.D.

Speech Communications Research Laboratory, Inc.

800A Miramonte Drive

Santa Barbara, CA 93109

Contract F44620-74-C-0034

Air Force Office of Scientific Research

Reporting Period: January 1, 1976 - December 31, 1976

April 18, 1977

A

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DDC
This technical report has been reviewed and is approved for public release IAW AFR 190-12 (7b).
Distribution is unlimited.
A. D. BLOSE
Technical Information Officer

The following work was accomplished during this contract year:

During the 1976 contract year our work for the Office of Scientific Research has accomplished the following goals:

(1) Implementation of programs and procedures for storing labels which contain descriptive information about speech events, thereby permitting sorting and retrieval of speech sounds based upon certain qualities or characteristics;

(2) The implementation of an automatic procedure for screening data for the purpose of locating potentially erroneous samples in a large data base;

(3) The formulation of an automatic boundary detection algorithm which is useful for locating the boundaries between speech sounds; and

(4) The formulation of an automatic steady-state detection algorithm which locates and labels the most stable interval of a speech sound.

SPEECH LABELS

The standard reference for the classification of speech events is the perceptual transcription. The coding of transcriptions into a computer has made it possible to tabulate statistics on natural speech for acoustic and linguistic studies. For making acoustic measurements on conversational speech which has been stored in computer files, a convenient method of referring to the acoustic segments is necessary, along with the transcriptional notation identifying each segment. This has been accomplished by means of a descriptive file which contains a label for each event of interest in the speech file. Programs can then access the label files for transcriptional information, as well as for the locations of the acoustic segments themselves. The label files therefore serve as directories of the speech files.

A label consists of two lines of text (ASCII) and it contains two basic kinds of information about a speech event, 1) its description, 2) its location. The description of the event is divided into the following six fields which make up the first line of the label.

1. Segment identification (20 characters) - contains some predefined set of symbols which identify the event being labeled.

2. Stress (two digits) - if the event has some stress level which can be quantified.

3. Environment (eight characters) - identification codes of the events adjacent to the labeled event can be stored. This field is symmetrically defined such that the four leftmost characters are for the left environments and the four rightmost characters are for the right environments.

4. Sequence number (five digits) - labeled events can be numbered if desired.

5. Word orthography (20 characters) - when labeling sub-word events it may be beneficial to retain the orthography of the word in which the event occurred.

6. Speaker initials (eight characters) - the initials of the speaker to whom the labeled event belongs.

The second line of a label contains information pertinent to the location of the event in the digitized speech file, plus some documentary items. There are six fields in the second line.

1. Starting sample point (nine digits) - sample point number of the initial boundary of the event.

2. Number points (nine digits) - number of sample points in the event.

3. Sampling frequency (five digits) - sampling frequency at which the event was digitized.

4. File name (24 characters) - name of the file in which the event is located. The complete specification also includes the device on which the file is located and the directory under which it is stored.

5. Date (nine characters) - date the label was created.

6. User initials (eight characters) - initials of the person doing the labeling.

The purpose of a label file is to act as a reference for the contents of a speech file. Labeling is performed at an interactive computer terminal where the researcher can use a time synchronized display of the acoustic wave, the RMS energy of the signal, and the formant frequency patterns to assist in locating the desired events. Once the segment boundaries have been marked

(either by hand or by an automatic process) the researcher is asked to enter the information for the first five fields of the first line of the label. All the remaining information in the label is supplied by the computer. The five descriptor fields are all optional so any number of them can be omitted.

As mentioned previously, the information content of any one label is descriptive enough that each label can stand on its own as an independent unit. Because of this, events can be labeled in any order and they can even be in different speech files. Likewise, labeling can be incomplete, i.e., if there is only interest in vowels then only the vowels need to be labeled.

DISTRIBUTION FILTERING

The collection of large scale speech data bases may require that the process of labeling relevant events in the acoustic signal be automated as much as possible. As a starting point for developing automated procedures we have investigated some characteristics of a manually labelled data base. This investigation was aimed at the following questions: 1) what type and magnitude of errors are present in acoustic measures on this data base, 2) what are the major sources of these errors, and 3) what can be done to reduce them.

The data base used for this study consists of 675 vowel tokens representing ten different vowel categories taken from a recorded interview with a single speaker. Vowels were labelled using a phonemic transcription. Transcription symbols were associated with the digitized acoustic waveform through an interactive process. The vowels were then analyzed for the frequency, bandwidth and amplitude of the first three formants by linear prediction analysis.

As a first step in characterizing the errors in acoustic measures on this data base, plots were made of the formant frequencies for all tokens of each vowel to observe trends in the data and gross excursions from these trends. While most data points had quite reasonable values for their respective vowels, there were several which departed radically from the central clusters. Possible sources of large errors which could explain the deviant samples are: 1) presence of inter-formant spectral

peaks, 2) merged spectral resonances, 3) labeling errors, 4) transcription errors, or 5) improper location of the steady-state analysis window. Regardless of the source of error, inclusion of strongly deviant samples would obviously degrade pattern recognition performance. Therefore, they must be detected and either corrected or excluded from consideration. Even if there is no gross error resulting from any of the five main sources, there will still be some random error associated with the numerical estimate of a correctly identified spectral peak. This simple measurement noise, however, appears to be small in comparison to actual fluctuations in the speech signal, and is therefore not a major concern in the present context.

Variability may be characterized by the mean and standard deviation of formant frequency distributions for each vowel category. The possible sources of variability can best be explored by examining individual vowel tokens whose parameter values lie outside of some range. For this purpose, we adopted a technique called distributional filtering which rejects tokens having values greater than a certain number of standard deviations from the mean. A similar technique was used by Peterson and Barney[1] to detect measurement errors.

A basic question with regard to these filters is, what is the optimum number of standard deviations which will include most

[1]Peterson, G. E. and Barney, H. L. (1952), Control methods used in a study of the vowels. J. Acoust. Soc. Amer., 24:2, 175-184.

of the good measurements and exclude most of the wild samples resulting from the above five sources of error? To explore this problem, we tested six different ranges from 0.5σ to 3.0σ. Plots were made of the percentage of the original number of tokens which passed the filter as a function of percentage expected for normally distributed, independent parameters. Only the vowels /i/, /ɪ/, /ɛ/ and /ə/ were used for this portion of the study because the greater number of tokens in these categories permitted more reliable statistics to be calculated for each filter condition. Despite differences in distributional characteristics, all four vowels showed similar trends, i.e., that filters with range less than about 2.0σ deviate markedly from expectation since far more tokens pass the filter than would be expected. This is probably attributable to the fact that the standard deviation of the original sample is a poor estimate for that of the underlying population with error points removed. We therefore chose 2σ as the range to study rejected tokens for sources of variability.

The 2σ filter rejected 75 out of a total of 501 tokens in the four vowel categories. When formant frequency measurements for the rejected tokens were examined in the context of surrounding sounds, it was found that 26 of these measurements could have been closer to population norms if the analysis window had been placed differently. This finding gives some indication of the need for more objective and perhaps automatic detection of the steady-state portions of vowels, since such an algorithm

would have greatly reduced the variability of dispersion and skew in the data base under study. Twenty of these 75 probably had incorrect transcriptions. It was observed that 23 of the remaining 29 tokens were heavily coarticulated with nasals and liquids, thus indicating that most of the residue of large departures are not attributable to gross errors, but indicate real variability in the speech signal.

AUTOMATIC BOUNDARY AND STEADY-STATE DETECTION

The results of the investigation just described suggested that automatic labeling procedures could help to eliminate errors in speech data bases. In studies of speech sounds the labeling process is typically done by hand, that is, the researcher uses a computer to analyze and display certain acoustic parameters and then a decision is made about the boundaries between sound units and about the location of the most steady-state portion of the sound (if applicable). In the performance of these tasks it is important to have a set of criteria or guidelines so that consistent decisions can be made. We have accomplished this by means of an algorithm which computes the spectral variance in the speech signal as a function of time. This variance function tends to have local maxima in the transition regions between sounds and local minima in sounds which can have steady-state characteristics. The appropriate maxima and minima are detected automatically and the resulting boundary and steady-state markers displayed for visual verification. The result is an elementary automatic segmentation process which is speaker independent and operates on conversational speech. This process gives us consistent algorithmic criteria for labeling speech sounds, thereby giving more reliable data samples. Furthermore, this represents the first step towards automatic data gathering, which is important collecting the large number of samples required in acoustic studies of speech sounds.

PUBLICATIONS

As a result of our research for AFOSR, several papers have been prepared:

A. Publications

1. June E. Shoup and Larry. L. Pfeifer, Acoustic Characteristics of Speech Sounds, Contemporary Issues in Experimental Phonetics, edited by Norman Lass, New York: Academic Press, pp. 171-224, 1976.

2. Ralph H. Fertig, Temporal Interrelations in Selected English CVC Utterances, SCRL Monograph No. 12, Speech Communications Research Laboratory, Inc., Santa Barbara, CA., 1976.

3. Larry L. Pfeifer, An Interactive Laboratory System for Research in Speech and Signal Processing, Submitted to IEEE Trans. on Acoustics, Speech and Signal Processing.

B. Presentation with Published Abstract

1. Robert J. Hanson and Larry L. Pfeifer, Labeling Speech Events for Acoustic and Linguistic Processing, presented at the 92nd meeting of the Acoustical Society of America, San Diego, CA., November, 1976. Abstract published in The Journal of the Acoustical Society of America, vol. 60, Supplement 1, pp. 512-513, 1976.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

(19) REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER (18) AFOSR - TR - 77 - 0631	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER (9)
4. TITLE *(and Subtitle)* (6) ACOUSTIC/LINGUISTIC ASPECTS OF AUTOMATIC SPEECH RECOGNITION	5. TYPE OF REPORT & PERIOD COVERED Interim / Rept. 1 Jan - 31 Dec 76
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) David J. / Broad, Larry L. / Pfeifer	8. CONTRACT OR GRANT NUMBER(s) (15) F44620-74-C-0034
9. PERFORMING ORGANIZATION NAME AND ADDRESS Speech Communications Research Lab, Inc. 800A Miramonte Dr. Santa Barbara, California 93109	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304/A2
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB DC 20332	12. REPORT DATE (11) 18 Apr 77
	13. NUMBER OF PAGES 10
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* (12) 13p.	15. SECURITY CLASS. *(of this report)* UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

(16) 2304 (17) A2

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

During the 1976 contract year the work for the AFOSR accomplished the following goals: (1) Implementation of programs and procedures for storing labels which contain descriptive information about speech events, thereby permitting sorting and retrieval of speech sounds based upon certain qualities or characteristics, (2) The implementation of an automatic procedure for screening data for the purpose of locating potentially erroneous samples in a large data base, (3) The formulation of an automatic boundary detection algorithm which is useful for locating the boundaries between speech sounds, and (4) The formulation of an automatic steady-state detection algorithm which locates and labels the most

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

290926

Block 20 – Abstract (continued)

stable interval of a speech sound.